

Blaž Cotman

Filozofska fakulteta Univerze v Ljubljani

Izziv »superintelligence« – komentar k Bostromovi knjigi *Superintelligence: Paths, Dangers, Strategies* (2014)

Članek orisuje zahteven izziv, ki ga predstavlja možnost »superintelligence«. Sledeč nekaterim ključnim poudarkom švedskega filozofa Nicka Bostroma, ki jih najdemo v njegovi knjigi *Superintelligence: Paths, Dangers, Strategies* (2014), skušam začrtati nekaj relevantnih filozofskih (metafizičnih, epistemoloških in etičnih) vprašanj, s katerimi smo soočeni, ko razmišljamo o umetnih agensih, ki presegajo človeško kognitivno zmogljivost na domala vseh pomembnih področjih. Začnem s kratko obnovo Bostromove »Nedokončane basni o vrabcih« in nadaljujem z opredelitvijo pojmov, kot so inteligenca, eksplozija intelligence in superinteligence. Zatem pojasnim, zakaj velja razlikovati med vprašanjem o tem, ali je neka entiteta inteligentna, in vprašanjem, ali poleg tega poseduje tudi zavedajoča se mentalna stanja, ter naznačim možne tehnološke poti do superintelligence in mnogovrstne negotovosti, ki obdajajo z njo povezana vprašanja. Trdovratnost in nerazrešenost teh vprašanj sugerirata, da preprostih rešitev preprosto ni in da imamo opraviti s perečim problemom, ki zahteva temeljito nadaljnjo filozofsko, znanstveno in nenazadnje tudi širšo družbeno pozornost.

Ključne besede: superinteligence, umetna inteligenca, zavest, Nick Bostrom

Uvod

V pričujočem članku bom predstavil nekaj pomembnih vprašanj in dilem, s katerimi nas v svoji novejši knjigi z naslovom *Superintelligence: Paths, Dangers, Strategies* sooča švedski filozof Nick Bostrom.¹ V njej pretresa kompleksne izzi-

¹ N. Bostrom (roj. 1973), nekakšen »znanstveni prerok« (Bostrom, 2009: 58), je direktor Oxfordskega inštituta za prihodnost človeštva, eden od ustanovnih članov Svetovne transhumanistične zveze, avtor knjig

ve, ki jih predstavlja možnost superinteligence, in na zvečine zgledno poglobljen način premišlja o tem, kako bi se lahko nanje najboljše odzvali (če in ko bi do »vzleta« superinteligence sploh prišlo, seveda, kar pa se zdi avtorju *zelo* verjetno, in sicer v ne tako daljni prihodnosti²). Po Bostromovem prepričanju je to »bržkone najpomembnejši in najstrašljivejši izziv, s katerim se je človeštvo kdajkoli soočilo« (Bostrom, 2014: vii).³

Ker gre za razmeroma obsežno, večplastno in zapleteno delo, ki ga na tem mestu niti slučajno ne morem zadovoljivo, kaj šele izčrpno, predočiti in analizirati, se bom v nadaljevanju omejil na oris tistih problematik, ki se mi zdijo še posebej relevantne in so iz takšnih ali drugačnih razlogov pritegnile mojo pozornost. Začel bom s kratko obnovo Bostromove »Nedokončane basni o vrabcih« (»The Unfinished Fable of the Sparrows«) in nadaljeval z opredelitvijo pojmov, kot so inteligenca, eksplozija inteligence in superinteligence. Zatem bom pojasnil, zakaj velja skrbno razlikovati med vprašanjem o tem, ali je neka entiteta inteligentna, in vprašanjem, ali *poleg tega* poseduje tudi zavedajoča se mentalna stanja, ter naznačil možne tehnološke poti do superinteligence in raznovrstne negotovosti, ki obdajajo z njo povezana vprašanja. V tem kontekstu bom med drugim skušal nakazati nepredvidljivost superinteligentnega vedenja ter izmuzljivost razmerja med moral(n)o(stjo) in inteligenco. Menim, da je razumevanje tega razmerja dokaj pomembno za prihodnost človeške vrste in planeta – zlasti v kolikor sprejmemo argumente v prid možnosti in/ali (precejšnji) verjetnosti eksplozije inteligence v (bolj ali manj bližnji) prihodnosti ter s tem idejo, da se bodo superinteligentni agensi prej ali slej pojavili, pa če to hočemo ali ne.⁴ Vse to se mi zdijo pomembni vidiki naslovne tematike, vredni temeljite nadaljnje filozofske, znanstvene in nenazadnje tudi širše družbene pozornosti.

Anthropic Bias: Observation Selection Effects in Science and Philosophy (2002) in zgoraj omenjene *Superintelligence: Paths, Dangers, Strategies* (2014), člankov, kot so »The Doomsday Argument is Alive and Kicking« (1999), »Are You Living in a Computer Simulation?« (2001), »Where Are They? Why I Hope the Search for Extraterrestrial Life Finds Nothing« (2008), »Cognitive Enhancement: Methods, Ethics, Regulatory Challenges« (2009), »The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents« (2012), »Existential Risk Prevention as Global Priority« (2013) in »Strategic Implications of Openness in AI Development« (2016), ter sourednik zbornikov *Human Enhancement* (2009) in *Global Catastrophic Risks* (2011).

² Bostromu se zdi *verjetno*, da se bo ključen preboj na področju umetne inteligence dogodil »enkrat v tem stoletju« (Bostrom, 2014: vii).

³ Bostrom pravi, da bo šlo za »bržčas zadnji izziv, s katerim se bomo spoprijeli« (Bostrom, 2014: vii). Zdi se namreč, da »bomo dobili eno samo priložnost« (prav tam) – ko bo superinteligence enkrat ustvarjena, bo naša usoda »zapečaten« (prav tam), v dobrem in/ali slabem.

⁴ Za koncizno predstavitev in ovrednotenje nekaterih temeljnih argumentov glej Chalmers, 2010.

Kratek pretres trdovratnega izziva superintelligence

Naj za začetek kar se da na kratko obnovim »Nedokončano basen o vrabcih«, s katero Bostrom prične svojo knjigo. Gre za skrajno sugestivno, poučno pripoved o vrabcih, ki v času napornega spletnja gnezd sanjarijo o sovi, ki bi jim pomagala pri tem duhamornem opravilu. Ker bi jih lahko znatno razbremenila dela, in sicer ne le tako, da bi jim pomagala pri že omenjeni gradnji gnezd, marveč tudi s tem, da bi pazila in skrbela za njihove najmlajše in najstarejše sorojake, oprezala za mačkami ter drugimi nevarnostmi ipd., se odločijo, da čimprej pričnejo z iskanjem sovinega jajca, izvaljeno ptico pa nato udomačijo in ustrezno izurijo. Do takšnega, na prvi pogled neproblematičnega načrta pa je že od samega začetka skeptičen enooki vrabec, ki v tem preprostem planu razpozna potencialno usodno nevarno pomanjkljivost – po jajce se odpravljajo, ne da bi pred tem razrešili problem udomačitve in urjenja sove, ko se bo ta enkrat izvalila iz najdenega jajca (Bostrom, 2014: v).

Četudi se vrabčji poglavar strinja, da je izziv, s katerim jih sooča dvomljivec, pomemben, nazadnje vendarle presodi, da je dovolj težavna že sama prisvojitve sovinega jajca in da je zato bolje, da se najprej lotijo kar te, o »problemu nadzora« oziroma (ne)možnosti(h) udomačitve in urjenja sove pa premlevajo in razpredajo pozneje. Razplet basni ni znan, Bostrom pa svojo knjigo posveti skeptičnemu vrabcu in njegovim somišljenikom, ki niso sodelovali pri »sovoiskateljstvu«, temveč so raje svarili pred nevarnostmi in skušali dognati odgovor na zagonetno vprašanje, kako sovo pravzaprav sploh vzgojiti, udomačiti in izuriti (prav tam). Zdi se mi, da se nekaj podobnega godi tudi z »iskanjem« (»močne«) umetne intelligence – večina proučevalcev hiti, da bi do nje kar najhitreje prispeli, le redki pa konkretnije in temeljiteje razmišljajo o tem, kakšne daljnosežne posledice bi utegnili imeti takšno prispetje in kako bi se veljalo že predhodno pripraviti na dogodek, ki bo, če bo do njega prišlo, po vsej verjetnosti »eden najbolj pomembnih dogodkov v zgodovini planeta« (Chalmers, 2010: 10).⁵

⁵ Na področju filmske umetnosti ne primanjkuje stvaritev, ki so upodobile dileme in probleme, ki jih predstavlja umetna inteligenca. Na tem mestu se bom omejil na omembo enega samega sodobnega primera, ki se mi zdi v okviru razpravljanja o umetni (super)inteligenci posebej relevanten in sugestiv. V *Ex Machini* (2015), najnovejšem filmu angleškega scenarista in režiserja Alexa Garlanda, uspe programerju Nathanu Batemanu (Oscar Isaac) ustvariti unikatno humanoidno robotko Avo (Alicia Vikander). Izkaže se, da premore ta – če sklepamo po njenem zavidljivo kompleksnem vedenju, vključno z njeno zmožnostjo koketiranja, pretvarjanja in nenazadnje premetenega načrtovanja poba – precej globlje in manj predvidljivo »notranje« življenje (ali, če hočete, simulacijo notranjega življenja), kot prvotno domnevata Nathan, ki ji pozneje celo izrecno pripiše zmožnost samozavedanja, ter njegov enotedenski pomočnik Caleb (Domhnall Gleeson). V kolikor bi se dejansko zgodilo kaj takega, kar prikazuje Garlandov film, bi to predstavljalo, če uporabimo Nathanove besede, »največji znanstveni dogodek v zgodovini človeštva«, ki bi lahko privedel do »eksplozije intelligence«, o kateri piše Bostrom: prvokrat bi bil ustvarjen »umetni« subjekt, »popolnoma etični ['umetni'] agens« (Tonkens, 2009), ki bi posedoval zmožnost samozavedanja, čustev, avtonomnega upora in predirljivega mišljenja, ki bi med drugim porajalo vprašanja takšne vrste, kot ga na neki točki zastavi Ava: »Mar ni čudno ustvariti nekaj, kar te sovraži?« Nemalo proučevalcev se strinja, da se je treba na vse pretege izogniti stvaritvi česa takšnega, kar bi nas

Bostrom se v nadaljevanju izogne in izrecno odreče rabi izraza »tehnološka singularnost«, ki je po njegovem mnenju zadobil pretirano »tehnoutopične konotacije« (Bostrom, 2014: 2).⁶ Prisega na »bolj precizno terminologijo« (prav tam), zaradi česar v kontekstih, v katerih bi običajno govorili o singularnosti, razpreda o »eksploziji intelligence« (ang. *intelligence explosion*), o kateri je slaven matematik I. J. Good (1916–2009) pisal že davnega leta 1965 v članku »Speculations Concerning the First Ultraintelligent Machine«. Temeljni pomen omenjene sintagme je koncizno pojasnil znani ameriški proučevalec umetne intelligence Eliezer Yudkowsky, ki je opozoril na to, da pri eksploziji intelligence ne gre za tezo o »splošni ekonomski in tehnološki pohitritvi« (Yudkowsky, 2013: 1), temveč je osrednja ideja v naslednjem: ker je tehnologija proizvod intelligence, bi v primeru stvaritve inteligentnih strojev slejkoprej prišlo do »učinka snežne kepe – prvi inteligentni umi bi lahko ustvarili inteligentnejše ume, inteligentnejši umi pa bi lahko izdelali še inteligentnejše ume« (prav tam) in tako naprej v potencialno neskončnost.

Toda kaj sploh sta inteligenca in superinteligenca? Inteligenco lahko razumemo kot »zmožnost doseganja ciljev v različnih okoljih« (Bostrom, 2014), kot »spособnost doseganja ciljev kljub oviram prek sprejemanja odločitev, ki temeljijo na racionalnih (resnici podrejenih) pravilih« (Pinker, 2009: 74).⁷ Superinteligenca pa

bilo *zmožno* (za)sovražiti in delovati v skladu s tem sentimentom. To pa razpira vprašanje, ali zamejitev zmožnosti, ki bi zadostovala za gotovo izognitev scenariju, ki smo mu priča v *Ex Machini*, ne bi obenem neizogibno pomenila onemogočenja stvaritve *resničnega* agensa in/ali *dejanske* superinteligence. Toda, kakor upravičeno opozarja Eliezer Yudkowsky, se s tem še vedno ne pritaknemo najglobljega in najpomembnejšega problema, ki ima opraviti s precej preprostejšo zmožnostjo (super)inteligentnih strojev – njihovo močjo, da krojijo svet in tako (so)oblikujejo (našo) prihodnost: »UI vas ne sovraži in prav tako ne ljubi, toda narejeni ste iz atomov, ki jih lahko uporabi za kaj drugega« (Yudkowsky, 2008: 333). Pri čemer gre tudi za to, da »[t]renutno zasedamo ogromno prostora na planetu«, in zatorej utegnemo predstavljati »potencialno grožnjo« (Agar, 2016). Več o tem v nadaljevanju.

⁶ Kot opozarja Marko Uršič, je futuristična raba izraza »singularnost« analoška in izvira iz matematike ter fizike: »V matematiki je singularnost točka, v kateri funkcija ni definirana (oz. ni odvedljiva). /.../ Singularnost v fiziki je 'črna luknja' v prostoru-času; pri tem pa je treba razlikovati med črno luknjo kot *območjem prostora-časa* za 'horizontom dogodkov' (kar pomeni, da informacije o dogodkih v njej niso dosegljive tistim, ki so/smo zunaj nje) – in samo singularnostjo, ki pa je zgolj teoretsko *hipotetična* točka v črni luknji oziroma znotraj območja prostora-časa za horizontom dogodkov, v kateri fizikalne količine domnevno presežejo vse končne vrednosti« (Uršič, 2015: 452). Tehnološka singularnost po besedah futurista Vernorja S. Vinge predstavlja točko, »v kateri moramo opustiti svoje stare [miselne] modele, saj zavlada nova realnost« (Vinge, v Uršič, 2015: 453), pri čemer je, kakor poudarja Uršič, »mišljena v dvojni analogiji s fiziko: gre za *enkratnost* tega Dogodka, za iztrganost Singularnosti iz zgodovinskega prostora-časa, in za njeno *zastrtost*, saj je (še) onkraj našega izkustvenega 'horizonta dogodkov'« (Uršič, 2015: 454).

⁷ Roman V. Yampolskiy, avtor nedavne knjige *Artificial Superintelligence*, inteligenco označi za »zmožnost najdejanja vzorcev v podatkih« (Yampolskiy, 2016: 98). Lakhmi C. Jain, Anas Quteishat in Chee Peng Lim pa so v članku »Intelligent Machines: An Introduction« opozorili, da so med »glavnimi dejavniki, vključenimi v "intelligenco", zmožnosti avtonomnega učenja in prilagajanja okolju« (Jain, Quteishat in Lim, 2007: 1). – Jasno je, da so vse navedene opredelitve zgolj provizorične in poleg tega ne implicirajo (nujno) druga druge, čeprav so medsebojno vsekakor tesno povezane. Vprašanje, katera definicija je najustreznejša, je zapleteno in ga na tem mestu puščam ob strani (za razčlenbo in zagovor – abstraktne, splošne – opredelitve intelligence, podobne zgornji(m), glej Legg, 2008: 1–22).

je po Bostromu pojmovana kot »razum, ki močno presega človeško kognitivno zmoglost na tako rekoč vseh pomembnih področjih« (Bostrom, 2014: 22).

Čeprav lahko v »vsakem diskontu /.../ kupimo računalnik, ki presega človeško sposobnost računanja, shranjevanja in priklica podatkov, risanja, preverjanja črkovanja, usmerjanja poti in tipografije« (Pinker, 2009: 95; prevod je nekoliko prirejen), se velja zavedati, da takšni računalniki še zdaleč niso superinteligentni, vsaj v Bostromovem smislu ne – njegova oznaka se namreč nanaša *zgolj in samo* na sisteme, ki imajo nadčloveško raven *splošne* intelligence.⁸

Naj na tem mestu izrecno pripomnim, da se zgoraj navedena definicija v ničemer vnaprej ne opredeli glede tega, za kakšno vrsto superintelligence naj bi šlo, temveč preprosto govori o razumu, ki izjemno presega človeško splošno kognitivno zmoglost. Prav tako ne pove ničesar o tem, ali se bo nadčloveško inteligentna entiteta dejansko zavedala svojega razuma, svojih presežnih »spoznavnih sposobnosti« in bo potemtakem svet izkušala prvoosebno, subjektivno, zavestno,⁹ tako da

⁸ Računalniki, s katerimi imamo opraviti dandanes, še vedno ne predstavljajo nečesa, kar bi »močno presevalo človeško kognitivno zmoglost na tako rekoč vseh pomembnih področjih« (Bostrom, 2014: 22). Se več, ne samo, da te zmoglosti ne presegajo – zaenkrat se človeški *splošni* inteligenci niti približajo ne, saj gre za »sisteme, ki imajo ozek nabor kognitivnih zmoglosti« (prav tam: 16). Bostrom se dobro zaveda, da so se »tehnične težave izdelovanja inteligentnih strojev izkazale za večje, kot so predvideli pionirji [UI]« (prav tam: 4), a poudarja, da to ne pove ničesar o tem, »kako velikanske so te težave in kako daleč smo trenutno od tega, da jih presežemo« (prav tam). Poleg tega ne smemo pozabiti, da se »[v]časih /.../ izkaže, da ima problem, ki sprva izgleda brezupno zapleten, presenetljivo preprosto rešitev (červeno je obratno verjetno pogostejše)« (prav tam).

⁹ S filozofom Richardom Brownom delim prepričanje, da gre pri vprašanju, »ali lahko dejansko ustvarimo sintetično ali umetno zavest«, za *odprto empirično* vprašanje, saj »se utegne izkazati, da je za zavest potrebno nekaj izrazito [*distinctly*] biološkega« (Brown, 2012: 212). To pomeni, da po mojem mnenju obstaja *možnost*, da je zavest biološki fenomen v smislu, da je »kavzalno vezana na določeno biokemijo svojih virov, kot so laktacija, fotosinteza ali katerikoli drugi biološki pojavi« (Searle, 1990: 379; prevod je nekoliko prirejen), in da je, denimo, bistveno odvisna od »fizikalnokemijskih lastnosti /.../ možganov« (prav tam), zaradi česar prav tako vnaprej ne izključujem možnosti, da je stvaritev »umetne«, nebiološke zavesti – ali »prenos, 'teleportacija' mentalnega softvera v neki *drugi*, ne nujno biološki hardver« (Uršič, 2015: 319) – nekaj nemogočega: »Nihče ne bo trdil, da lahko proizvajamo sladkor ali mleko tako, da z računalnikom simuliramo formalne sekvence laktacije oziroma fotosinteze, ko pa gre za vprašanje duha, je v takšne čudeže pripravljeno verjeti veliko ljudi, in sicer zaradi globoko zakoreninjenega dualizma; duh, menijo, je stvar formalnih procesov in je za razliko od mleka in sladkorja neodvisen od povsem določenih materialnih vzrokov« (Searle, 1990: 379; prevod je mestoma prirejen). Daniel C. Dennett izpostavlja in zagovarja verjetnost nasprotne možnosti, po kateri obstaja »(po)poln izomorfizem med človeško zavestjo in računalniškim programom/procesom« (Uršič, 1987: 197): »Če je jaz 'zgolj' pripovedno težišče in če lahko vse vidike človeške zavesti pojasnimo 'zgolj' z dejavnostmi virtualnega stroja, udejanjenega v izjemno prilagodljivih povezavah človeških možganov, potem bi ustrezno 'programiran' robot s silicijevimi računalniškimi možgani načeloma lahko bil zavesten in imel svoj jaz. Natančneje, lahko bi obstajal zavesten jaz z robotskim telesom in računalniškimi možgani« (Dennett, 2012: 509). Podobnega mnenja je tudi Douglas R. Hofstadter: »Duhovi obstajajo v možganih in bodo nemara obstajali tudi v programiranih strojih. Če in ko bodo takšni stroji nastali, njihove kavzalne moči ne bodo izvirale iz substanc, iz katerih bodo narejeni, temveč iz njihovega načrta in programov, ki bodo tekli v njih. Ali imajo te kavzalne moči, pa bomo zvedeli tako, da bomo govorili z njimi in skrbno prisluhnili tistemu, kar bodo imeli povedati« (Hofstadter, 1990: 389). A pripomniti velja, da celo iz tega, da je samozavedajoči se stroj nekaj nemogočega, *samo po sebi* (brez dodatnega argumenta) ne bi sledilo, da je nemogoča tudi stvaritev *superinteligentnega* stroja – vsaj ne, če sprejmemo (dozdevno sprejemljivo) tezo, da entitete, ki presegajo človeško (splošno) inteligenco, niso *nujno* tudi zavestne.

bomo lahko – à la Thomas Nagel – govorili in se spraševali o tem, *kako je biti* – »biti v 'prvi osebi'« (Nagel, 1990: 400) – ta entiteta.¹⁰ Bostrom se raje kot na vprašanja o metafiziki duha in/ali zavesti osredotoča na »vzročne antecedense in konsekvence superintelligence« (Bostrom, 2014: 22), s čimer se skuša izmakniti prenekaterim nadvse zagonetnim filozofskim problemom.¹¹ V članku »Why We Need Friendly AI?«, ki ga je napisal skupaj z Lukom Muehlhauserjem, razjasnjuje, čemu ubira takšen pristop: »Čeprav nekateri dvomijo, da bi lahko stroji posedovali mentalne lastnosti, kakršna je zavest, pomanjkanje tovrstnih lastnosti strojem ne preprečuje tega, da bi postali precej zmogljivejši od ljudi, da bi učinkovito krmarili prihodnost pri zasledovanju svojih ciljev« (Muehlhauser in Bostrom, 2014: 41). Uršič je podobno opozoril, da je vprašanje o tem, »ali ima lahko kompjuter (ali mreža ipd.) sploh svojo *lastno voljo*: ali lahko *avtonomno*, brez človeškega 'ukaza' sproži kakšno dejanje« (Uršič, 2015: 479), verjetno celo trdovratnejše in prvotnejše od vprašanja o zavestnosti superintelligence (vsaj z evolucijskega in, ne nazadnje, praktičnega ozira).¹² Tudi (super)inteligentni stroji, ki bi bili »zombiji«, oropani vseh subjektivnih doživetij, bi potemtakem predstavljali potencialno grožnjo naši blaginji.¹³ Kot pojasnjuje Rob Bensinger (2015), je te-

¹⁰ Kot pravi Nagel, ima »organizem [ali entiteta] zavedajoča se mentalna stanja tedaj in le tedaj, če obstaja nekaj takšnega kot *biti* ta organizem [ali entiteta] – nekaj takšnega *za* ta organizem [ali entiteto]« (Nagel, 1990: 400).

¹¹ Mednje – poleg vprašanja razlage intencionalnosti, subjektivno-kvalitativnega značaja izkustva, težkega problema zavestnega izkustva in brezštevilnih drugih izjemno zanimivih in težavnih znanstveno-filozofskih izzivov, ki so tako ali drugače povezani z duhom in zavestjo – sodi tudi vprašanje razmerja med inteligenco in (fenomenalno) zavestjo (ali – širše gledano – mentalnostjo). Naj pristavim, da so nekateri, med njimi zlasti prepričljivo Rosalind Picard, opozorili, da »imamo dokaze, da so čustva dejaven del intelligence, posebej zaznave, racionalnega mišljenja, odločanja, načrtovanja, ustvarjalnosti itn.« (Picard, 1997: 247), kar problem le še dodatno poglobi. Če namreč drži, kar trdi Picard, potem se *zdi*, da vprašanj (super)intelligence preprosto ni mogoče jasno razmejiti od vprašanj (super)čustev, posledično pa tudi ne od vprašanj (super)duševnosti.

¹² Uršič poudarja, da je na vprašanje avtonomnosti »nemara še težje odgovoriti kot na vprašanje, ali ima načeloma lahko kompjuter zavest, sámozavedanje. Kajti tudi če na vprašanje zavesti odgovorimo negativno (in osebno mislim, da kompjuterji, vsaj takšni, kakršne poznamo danes in kot si jih po vseh znanih scenarijih lahko zamislimo v bližnji prihodnosti, *ne* morejo imeti intencionalne zavesti, kaj šele sámozavedanja) – to še ne pomeni, da ne morejo biti avtonomni agensi« (Uršič, 2015: 479). Še več, takšne avtonomne, a ne(samo)zavedajoče se entitete niso zgolj možne, temveč tudi dejanske, saj jih – po vsem sodeč – najdemo v realnem svetu: živali so namreč »bitja, ki *so* avtonomni agensi, čeprav najbrž nimajo refleksivne zavesti (če izvajamo primata, morda). Pomislimo, recimo, na dinosavre: kako bi jim, če bi jih z biotehnoško genetiko znova oživili, vcepili kak 'etični kodeks', recimo, da ne bi ubijali človeških otrok, ki bi se gotovo radi igrali s temi pošastmi iz preteklosti v kakem realnem 'Jurskem parku'?« (prav tam).

¹³ Filozofi z izrazom »zombi« ne označujejo srh vzbujajočih »živih mrtvecev«, ki smo jih gledali in jih gledamo v številnih filmskih grozljivkah, marveč nekoga, »čigar vedenje je povsem naravno, pozorno, živahno in izrazno, a v resnici nima zavesti, temveč je nekakšen stroj« (Dennett, 2012: 97): »Glavna stvar pri filozofskem pojmu zombija je to, da med zombijem in normalno osebo ne moreš razlikovati samo na podlagi opazovanja zunanjega vedenja« (prav tam). Ali povedano natančneje: gre za *fenomenološke* zombije, za »bitja, ki so fizično istovetna ljudem, a so brez izkustev/doživetij: čeprav jih torej 'navzven' (na podlagi fiziološko-anatomskega ustroja in/ali vedenja) ni mogoče razlikovati od 'normalnih' ljudi, so 'navznoter' oropani vsakršnih fenomenoloških stanj« (Vörös, 2014: 266).

meljni vir Bostromovega zanimanja za inteligenco prav v »njemem potencialu, da rešuje praktične probleme in oblikuje prihodnost«, ¹⁴ kar je nekaj, kar bi nas moralo še posebej interesirati – in zaskrbljevati – z vidika varnosti. ¹⁵ Roboti, ki bi posedovali *avtonomno* zmožnost, da počnejo stvari, ki bi bile, denimo, uporabne za ljudi, bi, kakor so poleg Bostroma opozorili Colin Allen, Gary Varner in Jason Zinser (2000) ter mnogi drugi, imeli tudi zmožnost, da bi delali to, kar bi ljudem in drugim čutečim bitjem škodovalo. ¹⁶ To pa, kakor je dejala Rosalind Picard in si bomo natančneje ogledali v nadaljevanju, navsezadnje pomeni, da »večja kot bo svoboda stroja, bolj bo potreboval moralne standarde« (Picard, 1997: 134). ¹⁷

Možnih tehnoloških poti do superintelligence je več ¹⁸ (za nadrobnejšo razlago glej Bostrom, 2014: 22–51):

¹⁴ »Tudi če obstajajo druge vrste intelligence, so te domnevno manjšega ekonomskega in strateškega pomena kot 'kognitivne supermoči', ki jih Bostrom opisuje v 6. poglavju [knjige *Superintelligence*].« (Bensinger, 2015)

¹⁵ Ključen argument, ki ga podaja Bostrom, se da formulirati na način, ki bi se docela izognil rabi izraza »inteligenca« in bi ga, denimo, nadomestili »zmožnost programiranja« in »neka specifična sposobnost sklepanja« (Chalmers, 2010: 23). Nekaterim se zdi takšen argument celo prepričljivejši od Bostromovega (glej Bensinger, 2015).

¹⁶ Če nam uspe ustvariti umetno (splošno) inteligenco, potem bo to, kot že rečeno, nemara največja sprememba v človeški zgodovini in precej zlahka tudi njen konec. Ne pozabimo namreč, da bi šlo za inteligentne stroje, ki bi lahko reproducirali veliko večino dejavnosti naših možganov, pri čemer bi to delo celo opravljali mnogo hitreje in zanesljiveje od nas. Poleg tega bi lahko, kakor se bržkone razume samo po sebi, »za doseganje svojih ciljev /.../ uporabljali tudi spretnost prepričevanja, manipulacije in načrtovanja« (Bostrom, 2009). Če k temu dodamo še precejšnjo verjetnost »eksplozije intelligence«, ki bi sledila našemu stvarjenju umetnointeligentne entitete, potem bi se znalo zgoditi, da bi inteligenca strojev bliskovito tako zelo preseгла človeško, kot je človeška preseгла inteligenco hroščev ali črvov.

¹⁷ V kolikor bi imeli na neki točki kljub vsemu *dobre* razloge za verjetje, da smo uspeli ustvariti stroj, ki premore fenomenalno zavest, bi to porajalo nemalo zanimivih vprašanj. Ali bi bilo *moralno* sprejemljivo ugasniti ali uničiti stroj, ki bi bil zmožen zavedajočih se – ali kakršnihkoli – duševnih stanj (Dennett se npr. izizivalno sprašuje: »[Z]akaj naj bi bili razblinjeni upi 'zombija' manj pomembni od razblinjenih upov zavestne osebe?« (Dennett, 2012: 530))? Bi moral imeti takšen stroj pravice, kot so denimo pravica do življenja, svobode in varnosti? Bi ga bili »moralno obvezani priznati /.../ kot osebo in se posledično v odnosu do njega vzdržati vsega tistega, česar ne smemo storiti ljudem« (Harris, 2002: 29)? Bi bil v primeru, da bi bilo njegovo duševno življenje zadosti kompleksno, moralno odgovoren za svoje vedenje in bi si za svoje »moralne izbire« lahko resnično zaslužil grajo ali hvalo, nagrado ali kazen? Pripomniti velja, da se tudi ob vprašanju *dobrih* razlogov za prepričanje, da stroj poseduje zavest, zastavlja kopica vprašanj. Kakšni bi morali biti ti razlogi, da bi bili zadovoljivi? Bi za kaj takega *zadostovalo* dovolj sofisticirano – človeku podobno? – vedenje strojev? Če da, kdaj bi bilo določeno vedenje *dovolj* sofisticirano? In nenazadnje: mar ni možno (celo verjetno?), da bi robotovo (sámo)zavedanje vzniknilo, ne da bi to opazil kdorkoli izmed nas, ki robota (neizogibno) proučujemo zgolj »od zunaj«, brez vsakršnega dostopa do njegove notrine? Kdaj bi bilo, skratka, *upoštevajoč* vse te – filozofom razmeroma dolgo poznane – zagate, *racionalno* verjeti, da so stroji zavestni, in bi bilo do njih ustrezno privzeti to, kar Dennett imenuje »intencionalna drža« (ang. *intentional stance*), tako da bi »oddajalca zvoka /.../ obravnavaj[i] kot racionalni agens s prepričanj, željami in drugimi intencionalnimi ali 'naperjenimi' duševnimi stanji, katerega dejanja lahko razložimo (ali napovemo) na podlagi vsebine teh stanj« (Dennett, 2012: 101; moj poudarek)?

¹⁸ To dejstvo po Bostromovem mnenju že samo po sebi znatno poveča verjetnost prispetja na cilj po *najmanj* eni od njih.

- (1) eksponentno hiter (hardverski in softverski) razvoj računalnikov ter konstrukcija »močne« umetne (splošne) inteligence;
- (2) »emulacija« možganov, tj. stvaritev (super)intelligentnega softvera s pomočjo bolj ali manj filigranske preslikave/skeniranja »komputacijske« strukture bioloških možganov;
- (3) znatno izboljšanje delovanja bioloških možganov s pomočjo »selektivne vzreje« (ang. *selective breeding*) ali z biotehnološkim spreminjanjem genov (evgenika);
- (4) spajanje možganov/človeka in računalnika/stroja (nekakšna »kiborgizacija«);
- (5) obilen razvoj računalniških mrež in organizacij ter vznik »kolektivne superinteligence« ali celo »prebujenje« računalniških omrežij.

Bostrom meni, da sta najhitreje dosegljivi in hkrati najzanesljivejši poti močna umetna inteligenca in »emulacija« možganov, pripominja pa, da so šibke oblike superinteligence dosegljive tudi s pomočjo biotehnološkega nadgrajevanja kognicije, ki bi utegnili znatno pospešiti znanstveni in tehnološki napredek ter tako tudi pohitriti »vzlet« superintelligentne entitete ali celo množice njih. Manj obetavni se mu zdijo možgansko-računalniški vmesniki, ki bi po njegovem prepričanju v najboljšem primeru igrali podobno vlogo, kot jo igrajo »biološke poti« do nadčloveške inteligence.

Na tem mestu bi se rad, preden preidem k naznačitvi raznoraznih negotovosti, ki obdajajo vprašanje superinteligence, le še na kratko pomudil ob Bostromovem razpravljanju o »kinetiki« eksplozije inteligence. Pri tem gre za vprašanje hitrosti prehoda od stroja, ki bi dosegel človekovo raven kognitivnih sposobnosti, do superinteligence. Bostrom govori o treh možnih »tranzicijskih scenarijih«: (1) o počasnem scenariju, kjer govorimo o desetletjih ali stoletjih in ki pomeni idealno priložnost za relevantne družbenopolitične spremembe, za preizkušanje različnih pristopov, analiziranje situacije, skratka za temeljito pripravo na vznik superinteligence; (2) o zmernem scenariju, pri katerem gre za mesece ali leta (ter je časa za potrebno pripravo in spremembe nekoliko manj); (3) o hitrem scenariju, pri katerem govorimo o minutah, urah ali dnevih (kjer časa za poglobljene premisleke ter takšne in drugačne (proti)ukrepe tako rekoč ni). Vprašanje je torej, kako hitro po stvarjenju pametnega stroja, ki ga po intelektualnih sposobnostih ne bo več mogoče razlikovati od ljudi, bo prišlo do »vzleta« superinteligence kot »razuma, ki močno presega človeško splošno kognitivno zmožnost«. Bo šlo pri tem za minute, ure ali dneve? Za tedne? Ali pa celo za stoletja, tisočletja? Sloviti futurist Vernor Vinge je nekoč zapisal, da je najboljši odgovor na vprašanje, ali bodo računalniki kdajkoli tako pametni kot ljudje, najverjetneje sledeč: »Da, toda le začasno!« (Vinge, v Muehlhauser in Salamon, 2012: 15) Tudi Bostromu se zdi precej verjetno, da bo omenjeni prehod »eksploziven« in da se bo že stvarjenje strojne inteligence, ki bo po splošnih kognitivnih sposobnostih primerljiva z »vaškim beščkom« (Bostrom, 2014: 70), skoraj zagotovo izkazalo za težji in dolgotrajnejši iz-

ziv kot pa (softversko, hardversko in vsebinsko) izboljšanje tega sistema do te mere, da bo postal nadčloveško inteligenen.

Avtor nadalje upravičeno poudarja, da smo, ko premlevamo o vprašanjih, dilemah in problemih, povezanih s superinteligenco, »obkroženi z gosto tančico negotovosti« (prav tam: 255). Tisto, kar je pri vsem tem še posebej zaskrbljujoče, pa je, da je neznansko težko, če ne celo povsem nemogoče s kakršnokoli gotovostjo napovedati vedenje superinteligentne entitete. Lahko zgolj domnevamo, da bodo njena pota precej verjetno nenavadna, sploh gledano z naše razmeroma (ali če smo nekoliko bolj pesimistični, izredno) omejene perspektive. Razumevanje ozadja te ali one superinteligentne dejavnosti in ravnanja bo verjetno hudo presegalo meje naših umskih sposobnosti, in to *navkljub* temu, da je razumno predpostavljati, da obstajajo določeni instrumentalni cilji (npr. samoohranitev, kognitivna izboljšava, tehnološka izpopolnitev in pridobivanje virov), za katere si bo prizadeval tako rekoč sleherni inteligentni agens, vključno s superinteligentnim (glej Bostrom, 2012; Bostrom, 2014: 105–114).

A ne smemo pozabiti, da imamo v našem primeru bistveno prednost: *mi* bomo namreč tisti, ki bomo, če sploh bomo, stvaritelji »naših lastnih naslednikov«. To pomeni, da obstaja možnost, da določimo končne cilje superintelligence na ta način, da bo varovala človeške vrednote in delovala v skladu z veliko večino naših interesov. Toda kako doseči kaj takega? Pri tem nam nikakor ne more biti v pretirano uteho prepričanje o inherentni povezavi med inteligenco in moralnostjo, po katerem je superinteligentna entiteta po nekakšni nujnosti že tudi »supermoralna«. To po vsem sodeč ni nič drugega kot antropomorfní predsodek. Še več, kot opozarjata Bostrom in Muehlhauser, gre za »antropomorfizem«, ki v resnici ne drži niti za nas, saj »zlahka najdemo ljudi, ki so precej inteligentni, a nemoralni, ali takšne, ki so neinteligentni, toda popolnoma dostojni [*decent*]« (Bostrom in Muehlhauser, 2014: 44). Poleg tega se moramo zavedati, da »popolna vednost o moralni teoriji ne zagotavlja delovanja v skladu z njo« (Allen, Varner in Zinser, 2000: 260): »Lahko si zamislimo agensa, ki je docela zmožen razsojanja o tem, ali so določena dejanja moralna, a je hkrati povsem nemotiviran, da bi tudi deloval moralno. Takšen agens bo pogosto deloval na način, za katerega vé, da je [moralno] napačen. Pravzaprav je to opis nekaterih sociopatov.« (prav tam) Toda kakšno *zagotovilo* imamo, da superinteligentne entitete v tem oziru ne bodo podobne sociopatom? Zdi se, da prav nobenega.¹⁹

¹⁹ Bostrom poudarja, da bi nam bila lahko UI v svojih motivacijah precej manj podobna, kot bi nam bil, denimo, »zeleni luskasti Nezemljan« (Bostrom, 2014: 106), ki bi bil – predpostavimo – biološko bitje, ki se je, tako kakor mi, razvilo skozi evolucijski proces, tako da bi bilo upravičeno pričakovati, da bo imel »vrste motivacij, značilne za *razvita* [*evolved*] bitja« (prav tam; moj poudarek). Zato ni nič paradoksnega v možnosti, da bi bil edini končni cilj, ki bi ga utegnili imeti superinteligenca, »preštevanje zrnc peska na otoku Boracay« (prav tam: 107). Toda problem je v tem, da bi bil lahko celo tovrsten, navidez povsem nedolžen končni cilj *usoden* za človeštvo, saj bi imel preštevajoči agens v mnogih situacijah »konvergentni instrumentalni razlog«, »da pridobi neomejeno količino fizičnih virov in da, če je možno, odstrani

Kakor opozarja Rosalind Picard, bo sistem, ki deluje v tako kompleksnem in nepredvidljivem okolju, kakršno je naše, potreboval več kot zgolj takšne ali drugačne zakone (npr. tako ali drugače modificirane tri slavne Asimove zakone robotike) – »bistveno bo potreboval vrednote in načela, moralni kompas, nemara celo religijo« (Picard, 1997: 134). K temu velja pripomniti, da obstaja dober razlog za to, da *ne* ustvarimo nadčloveških strojev, ki bi z nami delili vse naše trenutne vrednote. Zakaj? Zamislimo si, da bi se že antični Grki soočili z »znanstveno-tehnološkim Prehodom«, o katerem govorimo v tem članku, in bi uspeli v stroj nekako vprogramirati svoje tedanje vrednote kot njegov končni cilj. »Z naše perspektive bi to,« kakor upravičeno poudarjata Bostrom in Muehlhauser (2014: 45), »rezultiralo v tragediji, saj menimo, da smo bili od antičnih Grkov dalje priča moralnemu napredku (npr. prepoved suženjstva).« Pri tem ne smemo spregledati, da smo najbrž tudi sami še vedno zelo daleč od kakršnekoli etične popolnosti (kaj je vendar bolj očitno od tega?), zaradi česar bi bilo superinteligenci pametno dopustiti nadaljnjo možnost moralnega napredovanja.

Kot vzpostavljevalci »začetnih pogojev« superinteligence se torej, na kratko rečeno, znajdemo pred kar najbolj kompleksnim *etično-filozofskim* vprašanjem – kakšen vrednostni sistem vprogramirati superinteligenci?²⁰ Ali z Uršičevimi besedami: »[K]atere cilje oz. smotre ali 'maksime' bi vpisali v 'začetne pogoje' superinteligence, če bi jih lahko še pred njenim prebujenjem (so)oblikovali?« (Uršič, 2015: 455). In nenazadnje: *kdo* naj o tem odloča? Jaz, »mi«, »vi«? Bodo to zagovorniki etike vrlin, deontologi, konsekvencialisti, kar vsi omenjeni ali kdo drug? Kako razrešiti brezštevilna nestrinjanja, s katerimi imamo, vsaj kar zadeva najzahtevnejše (filozofske) dileme, opraviti ne samo v filozofskih, temveč tudi vseh drugih logih? Menim, da Picard razpira ključen problem, s katerim se moramo soočiti, ko se sprašuje o tem, kdo bi imel *in* bi moral imeti »moralno avtoriteto nad računalniki, roboti, softverskimi agenci« (Picard, 1997: 134). Čeravno drži, da je takšna avtoriteta trenutno v rokah teh, ki razvijajo in programirajo računalnike (in/ali njihovih plačnikov), pa je to navsezadnje vsekakor »vprašanje za družbo kot celoto« (prav tam) – vprašanje, ki zadeva družbo *kot celoto* in ne le *nekaj* posameznikov znotraj nje.

potencialne nevarnosti« (prav tam: 116), vključno s človekom, ki sestoji »iz atomov, ki jih [superinteligenca] lahko uporabi za kaj drugega« (Yudkowsky, 2008: 333).

²⁰ Ryan Tonkens opozarja na še eno skrajno zanimivo, a zvečine spregledano razsežnost problema, ki sem se mu skušal približati v tem članku. Poudarja pomembnost vprašanja, ali bi bilo dani etični okvir – kakršenkoli bi ta že bil – sploh *zaželeno* in *moralno* implementirati v robote. Ker se zdi kantovski, deontološki, na dolžnosti(h) utemeljeni pristop mnogim najobetavnejši, se je Tonkens, da bi s tem spodbudil razpravo o tem, kar se mu zdi bistveno, osredinil nanj in pokazal, da bi bilo stvarjenje »kantovskih« umetnih agensov, ki bi izvajali etično relevantna dejanja, v resnici *antikantovsko* in – z vidika same etike, »vprogramirane« v te agense – *nemoralno* (za lucidno izpeljavo tega v marsičem presenetljivega sklepa glej Tonkens, 2009).

Zaključek

Če kaj, potem Bostromova knjiga vsekakor nazorno pokaže, da preprostih rešitev problema nadzorovanja superintelligence preprosto ni. Hkrati nas opominja na to, da je največkrat »mnogo lažje prepričati samega sebe, da si našel rešitev, kot pa jo dejansko najti« (Bostrom, 2014: 124). V tem oziru gre nedvomno za »pametno zamišljeno svarilo in napotek, da moramo začeti vsaj *premišljovati* o vprašanjih, povezanih z nadzorom vse sodobne vrhunske tehnologije, ne samo jedrske tehnologije ter biotehnologije, temveč tudi kompjuterske tehnologije ter informatike v najširšem pomenu« (Uršič, 2015: 480; moj poudarek).

Najbolj gotova stvar, ki jo po Bostromu lahko napovemo o prihodnosti, je, da se bodo dogajale spremembe. K temu pripominja, da kakor hitro »skušamo [te spremembe] natančneje opisati, takoj postanejo manj gotove« (Bostrom, 2009). Takšne misli dodobra odsevajo avtorjevo pogosto prav zavidljivo previdnost, s katero se loteva zapletenih tematik in problematik, ki jih obravnava. Tako na primer že uvodoma priznava, da je verjetno marsikatera njegova teza zmotna in poudarja, da »epistemska skromnost« ne zadostuje, temveč mora biti nadomeščena z nič manj kot »sistemskim priznanjem negotovosti in zmotljivosti« (Bostrom, 2014: viii). Lahko bi celo rekli, da gre pri Bostromu vseskozi za zavedanje tega, da je »edina futurološka napoved, ki je brez dvoma pravilna, /.../ da bodo v prihodnosti današnji futurologi videti trapasto« (Pinker, 2009: 96). Mar ni tovrstna – neprebrodjljiva – negotovost, s katero smo soočeni,²¹ že sama po sebi zadosten razlog za prebujenje iz brezskrbnega dremeža in za skrbno, poglobljeno premlevanje o superinteligenci, ki ostaja odprta možnost in bi nekoč, kot nas prepričljivo svari Bostrom, utegnila biti vzrok za znatno poslabšanje blaginje človeštva ali celo za eksistenčno katastrofo?²²

²¹ Kot pravi Dennett: »Nihče – vključno z mano – nima pregleda nad vsemi problemi in podrobnostmi, zato smo vsi v veliki meri zavezani k momljanju, ugibanju in pleteničenju.« (Dennett, 2012: 307)

²² Četudi se večina proučevalcev osredotoča na človeške živali in njihovo potencialno eksistenčno katastrofo (tj. izumrtje), ki bi bila posledica pojava superintelligence, v resnici seveda še zdaleč ne gre le zanje – v obzir je treba vzeti tudi trpljenje vseh ostalih čutečih bitij, ki bi ga utegnile povzročiti superinteligentne entitete (a ne nujno na tako, recimo temu, nepretanjene načine, kot smo ga do zdaj nemalokrat povzročali mi, pripadniki vrste *homo sapiens*).

The Challenge of "Superintelligence": A Comment on Bostrom's Book *Superintelligence: Paths, Dangers, Strategies* (2014)

The article outlines a series of complex challenges posed by the possibility of "superintelligence." In line with some of the key emphases made by the Swedish philosopher Nick Bostrom in his work *Superintelligence: Paths, Dangers, Strategies* (2014), I put forward some pressing philosophical (metaphysical, epistemological and ethical) questions that emerge when faced with the possibility of artificial agents surpassing human cognitive abilities in virtually all important areas of life. After a brief summary of Bostrom's "The Unfinished Fable of the Sparrows," I provide the preliminary definitions of the main concepts in the field of inquiry (i.e. intelligence, intelligence explosion, and superintelligence). I then explain why a distinction must be made between the question about the intelligence of an entity and the question about this entity also possessing conscious mental states, and go on to point out the possible technological paths towards superintelligence and the various uncertainties about related issues. The pressing nature of these questions and the fact that they still remain unresolved suggest that there are in fact no simple solutions, and that we are dealing with a serious problem that requires further thorough philosophical, scientific, as well as wider social engagement.

Keywords: superintelligence, artificial intelligence, consciousness, Nick Bostrom

Literatura

- Agar, N. (2016). »Don't Worry about Superintelligence«. *Journal of Evolution and Technology*, 26(1), str. 73–82. Dostopno na spletnem naslovu: <http://jetpress.org/v26.1/agar.htm>
- Allen, C., Varner, G. in Zinser, J. (2000). »Prolegomena to Any Future Artificial Moral Agent«. *Journal of Experimental & Theoretical Artificial Intelligence*, 12, str. 251–261.
- Bensinger, R. (2015). »Davis on AI Capability and Motivation«. Dostopno na spletnem naslovu: <https://intelligence.org/2015/02/06/davis-ai-capability-motivation/>
- Bostrom, N. (1998). »How Long Before Superintelligence?«. *International Journal of Future Studies*, 2, str. 1–13.
- Bostrom, N. (2003a). »Are You Living in a Computer Simulation?«. *Philosophical Quarterly*, 53(211), str. 243–255.
- Bostrom, N. (2003b). »When Machines Outsmart Humans«. *Futures*, 35(7), str. 759–764.
- Bostrom, N. (2005). »In Defense of Posthuman Dignity«. *Bioethics*, 19(3), str. 202–214.

Bostrom, N. (2009). »Znanstveni prerok«. *Mladina*, št. 4, str. 58–60. Dostopno na spletnem naslovu: <http://www.mladina.si/46052/znanstveni-prerok/>

Bostrom, N. (2012). »The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents«. *Minds and Machines*, 22(2), str. 71–85.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Brown, R. (2012). »The Brain and its States«. V Edelman, S., Fekete, T. in Zach, N. (ur.), *Being in Time: Dynamical Models of Phenomenal Experience*, Amsterdam: John Benjamins, str. 211–230.

Brundage, M. (2015). »Taking Superintelligence Seriously: Superintelligence: Paths, Dangers, Strategies by Nick Bostrom (Oxford University Press, 2014)«. *Futures*, 72, str. 32–35.

Chalmers, D. J. (2010). »The Singularity: A Philosophical Analysis«. *Journal of Consciousness Studies*, 17(9–10), str. 7–65.

Davis, E. (2015). »Ethical Guidelines for a Superintelligence«. *Artificial Intelligence*, 220, str. 121–124.

Dennett, D. C. (2012). *Pojasnjena zavest*. Ljubljana: Krtina.

Harris, J. (2002). *Vrednost življenja: Uvod v medicinsko etiko*. Ljubljana: Krtina.

Hofstadter, D. R. (1990). »Refleksije«. V Hofstadter, D. R. in Dennett, D. C. (ur.), *Oko duha: Fantazije in refleksije o jazu in duši*, Ljubljana: Mladinska knjiga, str. 380–389.

Jain, L. C., Quteishat, A. in Lim, C. P. (2007). »Intelligent Machines: An Introduction«. *Studies in Computational Intelligence*, 70, str. 1–9.

Legg, S. (2008). *Machine Super Intelligence (PhD dissertation)*. Lugano: University of Lugano. Dostopno na spletnem naslovu: http://www.vetta.org/documents/Machine_Super_Intelligence.pdf

Markič, O. (2011). *Kognitivna znanost: Filozofska vprašanja*. Maribor: Aristej.

Muehlhauser, L. in Helm, L. (2012). »Intelligence Explosion and Machine Ethics«. V Eden, A., Søraaker, J., Moor, J. H. in Steinhart, E. (ur.), *Singularity Hypotheses: A Scientific and Philosophical Assessment*, Berlin: Springer, str. 101–129.

Muehlhauser, L. in Salamon, A. (2012). »Intelligence Explosion: Evidence and Import«. V Eden, A., Søraaker, J., Moor, J. H. in Steinhart, E. (ur.), *Singularity Hypotheses: A Scientific and Philosophical Assessment*, Berlin: Springer, str. 15–40.

- Muehlhauser, L. in Bostrom, N. (2014). »Why We Need Friendly AI«. *Think*, 13(36), str. 41–47.
- Nagel, T. (1990). »Kako je biti netopir?«. V Hofstadter, D. R. in Dennett, D. C. (ur.), *Oko duha: Fantazije in refleksije o jazu in duši*, Ljubljana: Mladinska knjiga, str. 399–410.
- Picard, R. (1997). *Affective Computing*. Cambridge: MIT Press.
- Pigliucci, M. (2014). »Mind Uploading: A Philosophical Counter-Analysis«. V Blackford, R. in Broderick, D. (ur.), *Intelligence Unbound: The Future of Uploaded and Machine Minds*, Chichester, UK: Wiley-Blackwell, str. 119–130.
- Pinker, S. (2009). *Kako deluje um*. Tržič: Učila International.
- Searle, J. R. (1990). »Duhovi, možgani in programi«. V Hofstadter, D. R. in Dennett, D. C. (ur.), *Oko duha: Fantazije in refleksije o jazu in duši*, Ljubljana: Mladinska knjiga, str. 361–379.
- Tonkens, R. (2009). »A Challenge for Machine Ethics«. *Minds and Machines*, 19(3), str. 421–438.
- Uršič, M. (1987). *Matrice logosa: Filozofsko-logični eseji in študije*. Ljubljana: Državna založba Slovenije.
- Uršič, M. (2015). *Štirje časi: Filozofski pogovori in samogovori. Zima: Četrty čas, preludij: O sencah*. Ljubljana: Cankarjeva založba.
- Vörös, S. (2014). »Okrogel trikotnik ali miljo visok monocikel? O možnosti naturalizacije fenomenologije«. *Phainomena*, 23(90/91), str. 265–294.
- Yampolskiy, R. V. (2016). *Artificial Superintelligence: A Futuristic Approach*. Boca Raton: CRC Press.
- Yudkowsky, E. (2008). »Artificial Intelligence as a Positive and Negative Factor in Global Risk«. V Bostrom, N. in Čirković, M. M. (ur.), *Global Catastrophic Risks*, New York: Oxford University Press, str. 308–345.
- Yudkowsky, E. (2013). *Intelligence Explosion Microeconomics*. Berkeley, CA: Machine Intelligence Research Institute. Dostopno na spletnem naslovu: <https://intelligence.org/files/IEM.pdf>